


Model Fit and Item Factor Analysis: Overfactoring, Underfactoring, and a Program to Guide Interpretation

D. Angus Clark & Ryan P. Bowles


To cite this article: D. Angus Clark & Ryan P. Bowles (2018) Model Fit and Item Factor Analysis: Overfactoring, Underfactoring, and a Program to Guide Interpretation, Multivariate Behavioral Research, 53:4, 544-558, DOI: [10.1080/00273171.2018.1461058](https://doi.org/10.1080/00273171.2018.1461058)

To link to this article: <https://doi.org/10.1080/00273171.2018.1461058>

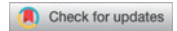
 View supplementary material 

 Published online: 23 Apr 2018.

 Submit your article to this journal 

 Article views: 292

 View Crossmark data 



Model Fit and Item Factor Analysis: Overfactoring, Underfactoring, and a Program to Guide Interpretation

D. Angus Clark and Ryan P. Bowles

Michigan State University

ABSTRACT

In exploratory item factor analysis (IFA), researchers may use model fit statistics and commonly invoked fit thresholds to help determine the dimensionality of an assessment. However, these indices and thresholds may mislead as they were developed in a confirmatory framework for models with continuous, not categorical, indicators. The present study used Monte Carlo simulation methods to investigate the ability of popular model fit statistics (chi-square, root mean square error of approximation, the comparative fit index, and the Tucker–Lewis index) and their standard cutoff values to detect the optimal number of latent dimensions underlying sets of dichotomous items. Models were fit to data generated from three-factor population structures that varied in factor loading magnitude, factor intercorrelation magnitude, number of indicators, and whether cross loadings or minor factors were included. The effectiveness of the thresholds varied across fit statistics, and was conditional on many features of the underlying model. Together, results suggest that conventional fit thresholds offer questionable utility in the context of IFA.

KEYWORDS

Categorical latent variable modeling; exploratory factor analysis; item factor analysis; model fit

Introduction

Exploratory factor analysis (EFA) is a widely used statistical technique for investigating the latent dimensional structure underlying a set of observed variables (Brown, 2013; MacCallum, 2009; Wirth & Edwards, 2007). EFA has many useful applications (Henson & Roberts, 2006; Kline, 1994; Russell, 2002), especially as a part of scale construction in the psychological and other social/behavioral sciences (Henson & Roberts, 2006; Kline, 1994). EFA summarizes lengthy assessments with a small number of factors, provides guidance to test users and administrators, can be used to generate scores on the latent dimensions of interest, and can provide important theoretical insights (Brown, 2013; Henson & Roberts, 2006; MacCallum, 2009). Furthermore, EFA often serves as a preliminary step for further analyses, such as more constrained measurement models, like confirmatory factor analytic models and Item Response Theory models, which typically require that the dimensional structure of the items being analyzed be known *a priori* (Brown, 2013; Henson & Roberts, 2006; Lee & Ashton, 2007; MacCallum, 2009; McDonald, 2013; Reise, Cook, & Moore, 2015). Thus, EFA is typically a key early stage in test development (Brown, 2013) when there is a large item pool (Lee & Ashton, 2007).

When performing an EFA, researchers must make a number of important decisions, with one of the most consequential being the number of latent dimensions to retain (i.e., the best characterization of item dimensionality; Preacher & MacCallum, 2003). Determining the optimal dimensionality of a set of items remains an imperfect art (Henson & Roberts, 2002), and though there are tools and heuristics to provide guidance, these methods cannot totally alleviate the common interpretational difficulties that arise (Goldberg & Velicer, 2006; Preacher & MacCallum, 2003). Notably, this issue is exacerbated in the context of item factor analysis (IFA), or EFA when indicators are categorical in nature (Wirth & Edwards, 2007). Traditional EFA, and the tools used to guide determinations of dimensionality, were developed for use with continuous data, and the application of these techniques to categorical data, especially dichotomous data, can lead to more suspect and difficult to interpret results (Ferrando & Lorenzo-Seva, 2000; Lee & Ashton, 2007; MacCallum, 2009).

Given the challenges associated with IFA relative to EFA with continuous outcomes, including added computational challenges (Wirth & Edwards, 2007), and the fact that traditional tools used to evaluate EFAs might be limited in this context (e.g., parallel analyses may become less

effective when scales contain dichotomous items; Yang & Xia, 2015), it is critical to consider any potential methods that could be used to aid researchers in making decisions regarding dimensionality in IFAs. As IFAs can be considered a special case of structural equation model (SEM), examining indices of model fit standard in SEM has been proposed as a method of helping to determine dimensionality in IFA (MacCallum, 2009). Model fit statistics are usually judged in reference to certain established thresholds, or “rules of thumb” (e.g., Hu & Bentler, 1999). However, these commonly invoked rules were derived from analyses with continuous data in a more confirmatory framework, without EFA/IFA in mind (West, Taylor, & Wu, 2012). Thus, the current study uses Monte Carlo simulation methods to examine the extent to which popular SEM model fit statistics and their commonly invoked cut-off values can help guide the determination of dimensionality in IFA with dichotomous indicators.

Item factor analysis

EFA was developed to identify common factors among normally distributed continuous variables; however, many assessments and questionnaires are made up of dichotomous or ordered categorical items (Wirth & Edwards, 2007). Although under certain conditions categorical variables can be treated as continuous without any major consequences (e.g., Rhemtulla, Brosseau-Liard, & Savalei, 2012), ignoring the categorical nature of items will often result in biased parameter estimates and fit statistics (e.g., Bandalos, 2014; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Rhemtulla et al., 2012). Treating dichotomous items—which represent the most extreme departure from continuity—as continuous generally leads to the most distorted conclusions (Rhemtulla et al., 2012). This is noteworthy as there are several fields and disciplines that rely heavily on dichotomously scored measures. For example, ability testing in educational settings, and symptom assessment in clinical psychology, are often based on questionnaires in which responses can fall into either one of two categories (i.e., correct/incorrect and symptom present/symptom absent).

Recent technological and statistical advances have made the EFA of categorical items (i.e., IFA) relatively accessible (Wirth & Edwards, 2007). Contemporary statistical programs are much more capable than their predecessors in bearing the relatively high computational burden associated with categorical latent variable modeling. Further, there have been many strides in the development of estimation techniques for categorical data. The two most popular categorical estimation approaches are maximum likelihood (ML), and mean and variance adjusted weighted least squares (WLSMV; Wirth

& Edwards, 2007). Each estimator represents a generally viable approach (Bandalos, 2014; Lei, 2009), though ML is more limited in that estimation quickly becomes intractable as the number of indicators and dimensions in an analysis increases (Wirth & Edwards, 2007). In the exploratory context of IFA, this disadvantage of ML can be particularly problematic, so WLSMV is generally the preferred estimation approach for IFAs.

Conceptually, the least squares approach to IFA is based on the assumption that underlying each categorical indicator is a normally distributed continuous latent response variable (Muthen, du Toit, & Spisic, 1997; Wirth & Edwards, 2007). An individual's standing on this latent variable relative to a set of thresholds determines which response category they fall into. For example, for a dichotomous item, if the individual's standing on the latent response variable is below a certain threshold they will endorse a score of 0, whereas if they are above this threshold they will endorse a score of 1. These continuous latent response variables and the correlations between them (tetrachoric or polychoric correlations depending on the number of response categories) are derived via ML estimation, and then the actual model parameters are estimated from these correlations using least squares estimation procedures. Most least squares estimators include a weight matrix, and the fit function of the traditional weighted least squares (WLS) estimator uses the inverse of the asymptotic covariance matrix of the polychoric correlations in this role (Muthen, 1993; Muthen et al., 1997; Shin, 2013; Wirth & Edwards, 2007). The estimation of this full weight matrix is difficult, however, and WLS will provide biased results unless sample sizes are exceptionally (often impractically) large (Dolan, 1994; Muthen et al., 1997; Wirth & Edwards, 2007). The WLSMV variation on WLS addresses this issue by using a diagonal weight matrix (as opposed to the full weight matrix) and adjusting standard errors and model test statistic for the fact that a nonoptimal weight matrix was used (Muthen, 1993; Muthen et al., 1997; Wirth & Edwards, 2007). WLSMV generally performs well at estimating parameter values and standard errors for IFA (Bandalos, 2014; Flora & Curran, 2004; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Lei, 2009; Muthen et al., 1997; Wirth & Edwards, 2007), though it may struggle under certain conditions, such as when there is a high amount of missingness in the data (DiStefano & Morgan, 2014; Savalei, 2011).

Although the computation of IFA models is no longer a major roadblock, these analyses still present considerable challenges when it comes to interpretation and selecting a robust and interpretable dimensional structure (Wirth & Edwards, 2007). Within traditional EFA, several tools with associated rules of thumb are typically used to help

determine dimensionality (Goldberg & Velicer, 2006). One of the most common can be referred to as the “K1” or “Kaiser” rule, in which as many factors are extracted as there are eigenvalues greater than 1 (Kaiser, 1960). Another popular method, examining the “scree plot”, entails looking for the first sharp “break” in a plot of the eigenvalues (Cattell, 1966). A more advanced technique, parallel analysis, involves comparing the actual eigenvalues to a plot of eigenvalues generated from random data with the same number of observations and items as the actual data (Horn, 1965).

Despite its popularity, the K1 rule is notoriously problematic and will often lead to overfactoring, that is, extracting too many factors (Preacher & MacCallum, 2003; Russell, 2002). Examining the scree plot can be helpful, but in practice it is often not obvious where the first major break is (Goldberg & Velicer, 2006; Lee & Ashton, 2007). Parallel analysis can also be effective, but is less accessible for models with categorical indicators (Russell, 2002), and has also demonstrated a tendency to overfactor in certain circumstances or underfactor (i.e., extracting too few factors) in others (Beauducel, 2001; Yang & Xia, 2015). The shortcomings of these traditional techniques, paired with the potential difficulty in generalizing them to IFA from EFA, leaves researchers with fewer tools at their disposal for guidance in determining the number of factors that best characterize their IFAs.

One potential diagnostic to guide evaluations of dimensionality is model fit (West et al., 2012). Indices of model fit have not been widely used in EFA, but represent an increasingly viable approach as SEM programs that can perform EFA/IFA become more widespread (MacCallum, 2009). SEM is predicated on attempting to reproduce an observed correlation/covariance matrix as closely as possible given a specified pattern of interrelations between variables. Indices of model fit are used to gauge how well the reproduced matrix matches the observed matrix, with different indices employing different approaches to quantify the mismatch. The same basic principle of reproducing an original data matrix underlies IFA, which can be considered just one specific instance of SEM (Brown, 2013). As such, fit statistics typical in SEM can potentially be brought to bear on the issue of determining dimensionality in IFA. Presumably, more appropriate factor structures will be better at reproducing the observed data, and this will be reflected in markers of model fit.

Model fit and item factor analysis

A plethora of indices and statistics for assessing model fit exist. The focus here will be on four of the most popular indices that are readily available for IFA with WLSMV estimation: the chi-square test of exact fit, the root mean

square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker–Lewis index (TLI) (West et al., 2012). The chi-square tests whether the model implied data matrix exactly matches the observed data matrix (after taking into account sampling error). The chi-square test applies the null-hypothesis significance-testing paradigm to the issue of model fit. If the test statistic is not significant, then it can be concluded that the reproduced matrix does not deviate from the observed matrix more than would be expected by chance; a significant value implies the opposite. The chi-square test was developed following the realization that when certain assumptions are met, and the model is correctly specified, the log likelihood values produced during the estimation procedure follow a central chi-square distribution. However, if any assumptions (e.g., multivariate normality) are violated, or the model is misspecified in any way, the log likelihood values follow a *noncentral* chi-square distribution. A consequence of this is that the power to detect any amount of misfit increases quickly with sample size, such that trivial amounts of misfit are likely to result in a significant chi-square (Browne & Cudeck, 1993). Given the untenability of the requirement that all models be perfectly specified and perfectly reproduce the observed data (MacCallum & Austin, 2000), many other fit statistics have been developed to complement the chi-square.

Although these other fit statistics place models on a continuum of fit, it is common practice to set benchmarks for “adequate” or “excellent” fit; currently, there is a standard set of thresholds (or “cutoff rules”) that get used in the literature (West et al., 2012). The RMSEA represents a weighted sum of the squared residuals (Browne & Cudeck, 1993; West et al., 2012). Lower RMSEA values denote better fit, and typically values below .08 are taken to indicate adequate fit, and below .05 are taken to indicate excellent fit (Browne & Cudeck, 1993; Hu & Bentler, 1999; West et al., 2012). The CFI quantifies the difference between the actual model and a null “baseline” model (i.e., a model where all variables are specified as orthogonal to each other) in reproducing the observed data. Higher values denote a greater improvement in fit over the baseline model, and typically values above .90 are taken to indicate adequate fit, and values above .95 are taken to indicate excellent fit (Hu & Bentler, 1999; West et al., 2012). The TLI is similar to the CFI, but it places a greater emphasis on parsimony (i.e., a model with more superfluous parameters will look better based on the CFI compared to the TLI). Like the CFI, values above .90 are taken to indicate adequate fit, and values above .95 are taken to indicate excellent fit (Hu & Bentler, 1999; West et al., 2012).

Most indices of model fit and their rule of thumb cutoff values were developed with models that exclusively address continuous data (Monroe & Cai, 2015). Further,

most of the work establishing and evaluating the standard thresholds for model fit are not based on EFA/IFA. Rather, this body of work tends to revolve around confirmatory factor analytic models (CFA; Brown, 2006). Although in one sense CFA is simply a restricted form of EFA (Brown, 2013), there are still important distinctions between these analyses (e.g., if rotation is used, the presence of a mean structure, and the ability to model residual correlations). Accordingly, the mostly CFA-based literature on model fit may provide some useful guidance on the topic of fit in EFA/IFA, but it could be problematic to haphazardly apply these indices and thresholds to EFA, especially in the even more distant categorical context of IFA (Monroe & Cai, 2015).

Notably, even in the context of continuous CFA, there is substantial evidence that fit indices and their cutoff values do not always work as intended. Fit indices themselves should ideally be sensitive only to model misspecification (and for certain indices, parsimony), and should be equally sensitive to misspecification across contexts. If this is not the case, then heuristic cutoff values have the potential to lead researchers astray. Although fit indices and cutoff values often can work as intended (e.g., Fan & Sivo, 2005), in addition to model misspecification, under certain circumstances fit indices appear to be sensitive to sample size (e.g., Beauducel & Wittmann, 2005), the amount of missing data present (e.g., Davey, 2005), non-normality (e.g., Yuan, 2005), model type (e.g., Fan & Sivo, 2007), the strength of factor loadings and factor covariances (e.g., Beauducel & Wittmann 2005; Davey, 2005), and the number of factor indicators/model size (e.g., Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998).

Importantly, one fit statistic—the RMSEA—has received considerable individual attention. The RMSEA is especially popular partly because under certain conditions it has a known approximate sampling distribution (a noncentral chi-square distribution), and so confidence intervals can be computed around the point estimates to provide a more fine grained assessment of fit (Browne & Cudeck, 1993; Curran, Bollen, Chen, Paxton, & Kirby, 2003). This foundational strength notwithstanding, specific investigations of the RMSEA's performance have revealed that it is also sensitive to factors such as sample size, nonnormality, the nature of misspecification (e.g., between vs. within factors), the type and size of the model, and the strength of factor loadings (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Kenny, Kaniskan, & McCoach, 2015; Nevitt & Hancock, 2000; Savalei, 2012).

In light of the above, there have been some calls to wholly abolish fit indices from the literature (e.g., Barrett, 2007). Others take a more reasoned approach, arguing that fit indices are an important part of model evaluation, but should be used prudently (e.g., Mulaik,

2007). In other words, the problem is not with the fit indices per se, but rather with the overapplication and universalization of heuristic “golden rules” of model fit (Markand, 2007; Marsh, Hau, & Wen, 2004; Saris, Satorra, & van der Veld, 2009; Steiger, 2007).¹ As such, the functionality of fit indices, and especially their typical cutoff values, ought to be evaluated under a variety of circumstances, so researchers can be familiar with the characteristics of their data and model that might result in the traditionally used cutoff values being too liberal or conservative.

Currently, it remains unsettled to what extent indices of model fit and their commonly adopted cutoff values can reasonably guide the evaluation of dimensionality in IFA models (Bovaird & Koziol, 2012; Edwards, Wirth, Houts, & Xi, 2012). Although fit indices have often been proposed as a method to guide the determination of optimal dimensionality in EFA (Hayashi, Bentler, & Yuan, 2007; Preacher, Zhang, Kim, & Mels, 2013), the practical utility of fit has not been widely evaluated even in the context of continuous EFA (Garrido, Abad, & Ponsoda, 2016). Most relevant investigations have focused on the chi-square test, considering either the first nonsignificant factor solution as the optimal number of dimensions or considering relative fit (i.e., comparing nested models and retaining the first factor solution that does not fit significantly worse, based on the chi-square test, than a solution with one or more extra factors), or the RMSEA with a .05 cutoff. Generally, exactly or approximately correct solutions will fit well. However, the chi-square test often tends toward overfactoring (e.g., Barendse, Oort, & Timmerman, 2015; Beauducel, 2001; Ferrando & Lorenzo-Seva, 2000; Frazier & Youngstrom, 2007; Hayashi et al., 2007). The RMSEA with its .05 cutoff has demonstrated usefulness at identifying the underlying factor structure of a set of indicators (Preacher et al., 2013), but as it involves the chi-square statistic, it may also suffer from some of the chi-square test's limitations (Hayashi et al., 2007). Similarly, there is some suggestion that the CFI/TLI may be better at identifying optimal factor structures without overfactoring as reliably as the chi-square test (Frazier & Youngstrom, 2007).

This literature suggests that considering model fit may be useful, at least when data are continuous, but there also may be some problems with certain indices and thresholds. Complimenting these results is research considering model fit with categorical data outside the IFA framework, which can inform the use of fit statistics in IFA. Typically, the chi-square does not appear to function too differently with categorical data using WLSMV estimation than its continuous ML counterpart

¹ It is worth noting for posterity's sake that in their seminal and widely cited paper regarding potential model fit guidelines, Hu and Bentler (1999) caution against the overapplication of their results.

(Bandalos, 2014), but comparative chi-square difference tests may have an inflated type one error rate (Sass, Schmitt, & Marsh, 2014). Under certain conditions the CFI and TLI appear capable of identifying properly specified models based on the .95 threshold, but both indices and cutoffs appear to become less stable as the number of categories decreases, and category asymmetry increases (Beauducel & Herzberg, 2006; DiStefano & Morgan, 2014). In some instances, the RMSEA appears relatively insensitive to category number and asymmetry (DiStefano & Morgan, 2014); however, there is also evidence that the RMSEA is in fact sensitive to these and other data features (Bandalos, 2008; Beauducel & Herzberg, 2006; Monroe & Cai, 2015). Monroe and Cai (2015) specifically concluded that the RMSEA cutoff values developed for continuous data are not suitable for categorical data.

Although these findings based on continuous EFA and estimation techniques for categorical data are informative, a handful of recent studies have more directly considered model fit as a tool for guiding dimensionality assessments in IFA. Barendse and colleagues (2015) examined the performance of the chi-square test and the RMSEA in models with both major and minor (i.e., “method”) factors. Their findings generally suggested that the RMSEA is better suited for identifying the major factor structure, whereas the chi-square test may be better for identifying the full model. Another important insight from this study was that comparative fit methods (i.e., comparing one factor solution to alternative models with more or fewer factors) tend to overfactor, often to the point that convergence becomes problematic. Consistent with what has been shown in continuous EFA, Yang and Xia (2015) showed that the chi-square test will often push toward overfactoring, as will the RMSEA and CFI when items are dichotomous. Garrido and colleagues (2016) considered several fit statistics at once and concluded that standard thresholds for model fit are not particularly well suited for identifying optimal factor structures in population models characterized by simple structure. However, they also noted that the CFI and TLI tended to perform the best in this regard.

In sum, there may be particular features of categorical data that undermine the utility of fit indices, and render the cutoff values developed for continuous data models inappropriate in the context of IFA. The limited (but growing) availability of evidence on this topic, and the utility of model fit for EFA generally, suggests both a potential usefulness for indices of fit in these contexts (e.g., Barendse et al., 2015), but also warrants caution (e.g., Garrido et al., 2016). Thus, there is a need to further evaluate the performance of fit indices and cutoff values in the context of IFA (Garrido et al., 2016).

Current study

The aim of the current Monte Carlo simulation study was to examine the utility of model fit indices and their frequently invoked cutoff values in helping to make decisions regarding dimensionality in IFA. That is, the goal was to demonstrate whether standard model fit thresholds can confidently be used to help identify the optimal number of latent variables underlying a set of categorical indicators. There are two major types of errors that can be made in this context: underfactoring and overfactoring (Kline, 1994; MacCallum, 2009). The former refers to identifying too few dimensions, and the latter, too many. Each error misrepresents the latent structure, and is more likely to provide difficult to interpret and unreproducible results. However, underfactoring tends to be more deleterious than overfactoring (Hayashi et al., 2007; Kline, 1994; MacCallum, 2009). Thus, we emphasize the ability of fit statistics and their cutoff values to prevent underfactoring.

Part of this emphasis is also conceptual, as the more factors that are included in an IFA solution, the better the reproduced data matrix will match the observed data matrix (Goldberg & Velicer, 2006). Thus, the nature of fit statistics is such that an overfactored solution will fit better than the correct/ideal solution. That is, underfactoring is more representative of the type of misspecification that indices of model fit have been developed to detect (i.e., they are largely sensitive to failures to reproduce the observed structure). Therefore, absolute fit thresholds will not be especially useful in discriminating between correctly and overfactored models. However, though models with more factors will always fit better, the cutoffs commonly applied in the literature may still reliably discriminate between underfactored and more optimally factored IFA models. Thus, to be useful in the context of IFA, fit cutoffs should generally reject underfactored models, while accepting correctly (and over) factored models.

In this study we specifically examined the performance of several popular model fit indices (chi-square, CFI, TLI, and RMSEA), and the recommendations of their commonly applied cutoff values, when dichotomously scored assessments of various length were subjected to IFA. We focus on dichotomous items because of their prevalence in testing and other data settings, and the evidence that fit indices and cutoff values may be least trustworthy with fewer categories. Fit indices were examined under several population model conditions, and with varying degrees of misspecification; we varied the number of items in the model, the strength of the factor loadings, and the strength of the intercorrelations between factors. All of these variables have been shown to relate to model fit before in either continuous or categorical contexts.

Indicators also varied in their distributions, ranging from symmetrical to asymmetrical, because in practical applications it is likely that there will be a mix of symmetrical and asymmetrical items, and there is also evidence that category asymmetry can affect the evaluation of model fit.

With these simulations we hope to build on and extend the small but growing body of work on model fit in the context of IFA (e.g., Barendse et al., 2015; Garrido et al., 2016). Here, we contribute to this existing literature by considering several popular indices of fit, longer assessments, and multiple features of the underlying population model, including model complexity with major and minor factor structures. All analyses in this study were based on the popular and powerful Mplus program (Muthén & Muthén, 1998–2015), its WLSMV estimator, and its approach to computing fit indices. We also provide a program that can be used by researchers to help guide the interpretation of fit indices when working with dichotomous items. It is our hope that this program, paired with the results from this and previous studies, will facilitate more effective use of fit statistics when exploring the underlying factor structure of questionnaires and assessments.

Method

Data generation

Data were generated using the Monte Carlo feature of Mplus version 7.4 (Muthén & Muthén, 1998–2015). Initially, analogous continuous data and categorical data population models were specified. Given the relative lack of research on the performance of fit indices and their cutoffs in the context of EFA in general, the continuous models were used to establish a baseline from which to evaluate the categorical models that were of primary interest.

Population models followed a three-factor structure in which one third of the indicators loaded on each factor (i.e., each item loaded on one factor, and all other loadings were set to 0). All items within a given condition loaded on their respective factor to the same degree. Residual variances were set such that the total variance of each indicator was unity (i.e., residual variance = $1 - \lambda^2$). Correlations were specified between all three factors. Inter-factor correlations were constant in size within a condition. Item thresholds (or intercepts in the continuous population models) were selected to fall between the values of -2 and 2, ensuring that there would be a mix of symmetrical and asymmetrical items in each condition.²

² Eight-item thresholds fell between 0 and .50; six-item thresholds fell between .51 and 1.00; two-item thresholds fell between 1.01 and 1.50; and four-item thresholds fell between 1.51 and 2.00.

Threshold values varied within but not between factors (e.g., the first and second items in factor 1 had different thresholds, and the first and second items in factor 2 had the same two thresholds as the first two items on factor 1), and remained constant across population models. Factor means were set to 0, and factor variances were set to 1. All data sets were generated with 500 observations, a reasonable sample size for factor analysis in the psychological and other behavioral sciences. For every condition, 1000 unique data sets were generated.

Three properties of the population models were systematically varied: the size of the factor loadings, the size of the factor intercorrelations, and the number of indicators. Factor loadings were either high (.70), moderate (.50), or low (.35). Factor intercorrelations were likewise either high (.70), moderate (.45), or low (.20). The assessment length was either short (15 total items; 5 items per factor), medium (30 total items; 10 items per factor), or long (60 total items; 20 items per factor). Four different factor solutions were considered for each population model, a three-factor solution (i.e., the correct model), a two-factor solution, a one-factor solution, and a four-factor solution.

Data generation—follow-up analyses

Population models characterized by simple structure provide a “best-case scenario” for evaluating fit statistic performance, but may not be representative of the data researchers typically encounter. Thus, the generalizability of the trends observed in the initial analyses, which were based on simple population factor structures in which each item loaded on only one factor to a uniform degree, was briefly probed by considering two types of alternative population models that demonstrated more complexity. The follow-up models were extensions of the previous models, specified and evaluated using the same general procedure described for the initial analyses. For simplicity, we focused on models with 60 items.

First, we included cross loadings in which the last five items of every factor (i.e., $\frac{1}{4}$ of each factor's items) also loaded onto the next subsequent factor (e.g., items 15–20 had major loadings on factor 1, and minor loadings on factor 2). These cross loadings were set to have a magnitude of .25 across all conditions. All other items were specified to load on every other factor with a loading of .10. Residual variance values were adjusted to ensure that the total variance of each item remained 1.

Second, consistent with Barendse and colleagues (2015), population models with minor factors (i.e., extra factors that only a small number of items cross-load on to) were considered. In addition to the three major factors of interest in these population models, five minor

factors were also specified. Three items, one from each major factor, loaded on to each of the minor factors. Loadings on the minor factors were set to .25. There were no other cross loadings in this model. All minor factors were orthogonal to each other, and to the major factors. Again, residual variance values were adjusted to ensure that the total variance of each item remained 1.

Data analytic strategy

For each individual replication, the model in Mplus was specified as an exploratory structural equation model (ESEM; Asparouhov & Muthén, 2009) with an oblique geomin rotation,³ estimated via WLSMV. One to four factors were specified, each with a mean of 0 and variance of 1. Every item was specified to load on every factor and every item and factor was specified to be part of the same exploratory structure (all items can potentially load on all factors included in the exploratory structure). With these specifications, the ESEM model is equivalent to an IFA with the specified number of factors. The ESEM specification was used here instead of the more straightforward EFA/IFA specification because the Mplus output files for ESEM models are more compatible with the program used to extract and integrate fit information across replications. Individual Mplus input files were generated and run for each condition using the Mplus Automation Package (Hallquist & Wiley, 2015) in R (R Core Team, 2016).

The Mplus Automation package was also used to extract fit information from the individual output files. For each condition, there were 1000 total output files. Four statistics for each of the fit indices of interest were computed from the information included in the 1000 output files per condition. For the chi-square, RMSEA, CFI, and TLI the mean and standard deviation across all models within a condition was computed. For the chi-square test, the proportion of models that demonstrated statistically significant misfit at both the .05 and .01 alpha level was computed. Similarly, for the RMSEA, the proportion of models with values below .08 and .05 was computed. Finally, for both CFI and TLI, the proportion of models with values above .90 and .95 was computed. Importantly, the number of models that failed to converge for a condition was also calculated, and the proportion of models evidencing “acceptable” and “excellent” fit were calculated excluding the models that did not converge. Models with so-called “inadmissible solutions” or “Heywood Cases” (negative residual variances; Wothke, 1993) were not excluded as here these outcomes just represent

chance sampling error (as opposed to model error), and should not affect fit (Briggs & MacCallum, 2003; Marsh et al., 1998).

Results

The initial results are presented by fit statistic. First, the chi-square is discussed, followed by the RMSEA, and the CFI and TLI. Results from the continuous data models are reviewed first to establish a point of reference for the categorical models of primary interest. For ease of presentation, the subsequent sections and tables primarily focus on the 60-item condition. Most patterns described emerged across the 60-, 30-, and 15-item conditions. Any differences across assessment length are noted in the text. Full results are available online in the supplemental materials located on the open science framework (osf.io/f2xua).

Across all conditions, the majority of replications converged. Consistent with previous research (e.g., Barendse et al., 2015; Garrido et al., 2016), almost all replications that failed to converge were overfactored four-factor models. The four-factor models were less likely to converge as the number of items decreased, likely because there was less information available to capture a spurious fourth factor, leading to unstable estimation. The number of nonconverged four-factor solutions (categorical) ranged from 0 to 3 in the 60-item conditions, 9 to 32 in the 30-item conditions, and 121 to 429 in the 15-item conditions. A similar, albeit greatly attenuated, trend was observed for the other factor solutions; nonconvergence was rare overall, but more likely in the 15-item condition. Again, estimation may have been more unstable when there was less information for the models to draw in separating out multiple factors.

Chi-square test

When data were continuous, the chi-square test indicated statistically significant misfit for almost all underfactored models. Only when factor loadings were low and factor intercorrelations were high was the two-factor model not rejected in every replication. The three and four factor models demonstrated much less statistically significant misfit than the underfactored models, but more than would be expected given the *p* value and appropriateness of the model. However, across conditions the rate of type I error was stable, with roughly 40% of three-factor models consistently demonstrating statistically significant misfit at the .05 level (20% at the .01 level), and roughly 15% of four-factor models demonstrating statistically significant misfit at the .05 level (5% at the .01 level). Thus, when data were continuous the chi-square test reliably rejected

³ The default in Mplus in EFA-style analyses is to apply an oblique geomin rotation. The type of rotation, or whether there even is a rotation, does not have any ramifications for model fit or the results presented below (Brown, 2013).

Table 1. Fit results for the chi-square test across study conditions for 60-item assessment with dichotomous items.

	High loadings			Moderate loadings			Low loadings		
	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>
Mean									
One factor	2165.39	2981.84	4148.35	1877.39	2088.84	2416.69	1788.11	1839.36	1914.33
Two factor	1911.98	2273.84	2845.10	1745.88	1847.78	2001.24	1688.59	1720.20	1758.55
Three factor	1697.37	1694.15	1683.38	1638.03	1647.25	1649.19	1607.28	1623.72	1630.79
Four factor	1566.67	1573.13	1573.73	1552.07	1553.77	1553.03	1532.36	1543.16	1546.60
Standard deviation									
One factor	86.14	150.04	239.48	62.71	61.95	92.65	48.28	59.64	50.84
Two factor	67.80	103.32	148.53	34.88	48.08	64.64	27.60	31.48	36.13
Three factor	50.15	69.73	63.86	28.25	36.47	40.06	24.66	26.54	29.78
Four factor	26.37	26.77	25.14	23.97	24.94	24.90	23.14	23.90	24.71
<i>p</i> < .05									
One factor	100%	100%	100%	96%	100%	100%	25%	75%	100%
Two factor	100%	100%	100%	46%	99%	100%	3%	19%	59%
Three factor	52%	45%	36%	6%	13%	14%	0.10%	1%	4%
Four factor	2%	3%	2%	0%	0%	0%	0%	0.10%	0%
<i>p</i> < .01									
One factor	100%	100%	100%	69%	100%	100%	8%	32%	95%
Two factor	99%	100%	100%	12%	91%	100%	0.10%	3%	20%
Three factor	23%	19%	15%	0.10%	3%	5%	0%	0.10%	1%
Four factor	0%	0%	0%	0%	0%	0%	0%	0.10%	0%

Note. High *r*, high factor intercorrelations; Moderate *r*, moderate factor intercorrelations; Low *r*, low factor intercorrelations; *p* < .05, percentage of replications with a chi-square *p* value below .05; *p* < .01, percentage of replications with a chi-square *p* value below .01.

underfactored models, though it did demonstrate an elevated, yet consistent, type I error rate for correctly and overfactored models.⁴

Results were less consistent across conditions when considering the categorical models (see Table 1).⁵ Underfactored models, typically evidenced statistically significant misfit when factor loadings were high or moderate, but were less consistently rejected when factor loadings were low, and when factor intercorrelations were stronger. Correctly specified models in the 60-item conditions often evidenced more statistically significant misfit than expected given the alpha level when factor loadings were high or moderate (the rate of statistical significance reached as high as 52%), but not when factor loadings were low. The strength of the factor intercorrelations was inconsistently related to three-factor models evidencing statistically significant misfit; when factor loadings were high, statistically significant misfit was most common when factor intercorrelations were also high, but the opposite pattern emerged when factor loadings were moderate. Notably, there were fewer instances of statistically significant misfit in three-factor models at any level when there were fewer items (see: osf.io/f2xua). When there were 30 items, rates of statistical significance were either in line with, or slightly lower than, the stated .05 and .01 alpha levels. When there were 15 items, the chi-square was never significant more than 1% of the time. Overfactored models almost never evidenced statistically

significant misfit at either the .05 or .01 level; rates of rejection for overfactored models tended to be below the stated alpha level.

RMSEA

The thresholds used here were values below .08, and below .05. When data were continuous, most models appeared to fit adequately based on the .08 threshold. The .05 threshold rejected most underfactored models when loadings were high, but in the other conditions all models continued to evidence acceptable fit. Overall, when data were continuous the RMSEA suggested that most models, regardless of under or overfactoring, fit the data adequately (<.08), if not excellently (<.05).

These trends were more pronounced when considering the categorical models (see Table 2). All underfactored solutions across all conditions evidenced at least adequate (<.08) fit to the data, and the vast majority also demonstrated excellent (<.05) fit. The one exception was that 1 factor solutions were rejected at the .05 level when factor loadings were high, and intercorrelations were low. Notably, underfactored models were slightly more likely to be rejected at the .05 level in the 30- and 15-item conditions, but only when factor loadings were high. Consistent with these trends, correctly specified (i.e., three-factor) models, and overfactored models, were never rejected.

CFI/TLI

The CFI and TLI are presented together because of their conceptual similarity, and the fact that the results for each index were analogous (though on average the TLI was slightly more likely to reject models than the CFI). When

⁴ Complete results for the continuous model analyses can be found in the online supplemental material (osf.io/f2xua).

⁵ Figures that graphically depict results across the fit statistics and model types can be found in the online supplemental material (osf.io/f2xua).

Table 2. Fit results for RMSEA across study conditions for 60-item assessment with dichotomous items.

	High Loadings			Moderate Loadings			Low Loadings		
	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>
Mean									
One factor	.02	.04	.05	.01	.02	.03	.01	.01	.02
Two factor	.02	.03	.04	.01	.02	.02	.01	.01	.01
Three factor	.01	.01	.01	.01	.01	.01	.00	.01	.01
Four factor	.01	.01	.01	.00	.00	.00	.00	.00	.00
Standard deviation									
One factor	.002	.002	.003	.002	.002	.002	.003	.002	.002
Two factor	.002	.002	.002	.002	.002	.002	.003	.002	.002
Three factor	.003	.003	.003	.003	.003	.003	.003	.003	.003
Four factor	.003	.003	.003	.003	.003	.003	.003	.003	.003
<.08									
One factor	100%	100%	100%	100%	100%	100%	100%	100%	100%
Two factor	100%	100%	100%	100%	100%	100%	100%	100%	100%
Three factor	100%	100%	100%	100%	100%	100%	100%	100%	100%
Four factor	100%	100%	100%	100%	100%	100%	100%	100%	100%
<.05									
One factor	100%	100%	5%	100%	100%	100%	100%	100%	95%
Two factor	100%	100%	100%	100%	100%	100%	100%	100%	100%
Three factor	100%	100%	100%	100%	100%	100%	100%	100%	100%
Four factor	100%	100%	100%	100%	100%	100%	100%	100%	100%

Note. High *r*, high factor intercorrelations; Moderate *r*, moderate factor intercorrelations; Low *r*, low factor intercorrelations; < .08, percentage of replications with a RMSEA value below .08; < .05, percentage of replications with a RMSEA value below .05.

data were continuous, underfactored models tended to be rejected using the .90 threshold unless factor intercorrelations were high, in which case two factor models were likely to demonstrate adequate fit. When the .95 threshold was used, the majority of underfactored models were rejected. However, based on the .95 threshold, three- and four-factor models were more likely to be rejected when factor loadings were smaller. Thus, when data were continuous, the CFI and TLI generally rejected underfactored models and accepted correctly and overfactored models, especially when the .95 threshold was used.

When data were categorical the performance of these statistics was less consistent across conditions, but the patterns observed often represented an extension of those observed with continuous data (see Tables 3 and 4). Underfactored models still tended to be rejected at both the .90 and .95 level. However, underfactored models were even less likely to be rejected as the strength of the factor intercorrelations increased. When factor intercorrelations were high, one- and two-factor models were considerably more likely to appear acceptable at both the .90 and .95 level; even when factor intercorrelations were

Table 3. Fit results for CFI across study conditions for 60-item assessment with dichotomous items.

	High loadings			Moderate loadings			Low loadings		
	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>
Mean									
One factor	.94	.80	.55	.94	.81	.55	.89	.76	.52
Two factor	.97	.90	.78	.96	.90	.77	.95	.87	.75
Three factor	.99	.98	.98	.98	.97	.96	.98	.94	.91
Four factor	.99	.99	.99	.99	.99	.99	.99	.98	.96
Standard deviation									
One factor	.02	.03	.04	.02	.04	.05	.06	.08	.08
Two factor	.01	.02	.03	.02	.03	.04	.04	.06	.07
Three factor	.01	.01	.01	.01	.02	.03	.03	.04	.06
Four factor	.00	.00	.01	.01	.01	.01	.02	.03	.04
>.90									
One factor	99%	0%	0%	94%	0%	0%	53%	1%	0%
Two factor	100%	58%	0%	100%	50%	0%	88%	32%	1%
Three factor	100%	100%	100%	100%	99%	97%	99%	81%	59%
Four factor	100%	100%	100%	100%	100%	100%	100%	98%	92%
>.95									
One factor	28%	0%	0%	26%	0%	0%	12%	0%	0%
Two factor	91%	0%	0%	83%	1%	0%	51%	6%	0%
Three factor	100%	98%	98%	99%	86%	75%	81%	45%	29%
Four factor	100%	100%	100%	100%	100%	98%	95%	80%	67%

Note. High *r*, high factor intercorrelations; Moderate *r*, moderate factor intercorrelations; Low *r*, low factor intercorrelations; > .90, percentage of replications with a CFI value above .90; > .95, percentage of replications with a CFI value above .95.

Table 4. Fit results for TLI across study conditions for 60-item assessment with dichotomous items.

	High loadings			Moderate loadings			Low loadings		
	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>	High <i>r</i>	Moderate <i>r</i>	Low <i>r</i>
Mean									
One factor	.94	.79	.53	.93	.80	.53	.89	.75	.50
Two factor	.96	.90	.77	.96	.89	.76	.94	.86	.73
Three factor	.99	.98	.98	.98	.97	.96	.98	.94	.90
Four factor	.99	.99	.99	.99	.99	.99	1.00	.99	.97
Standard deviation									
One factor	.02	.03	.04	.03	.04	.06	.06	.08	.08
Two factor	.01	.02	.03	.02	.03	.04	.04	.06	.08
Three factor	.01	.01	.01	.01	.02	.03	.04	.05	.07
Four factor	.00	.01	.01	.01	.02	.02	.04	.05	.07
> .90									
One factor	98%	0%	0%	92%	0%	0%	50%	1%	0%
Two factor	100%	42%	0%	99%	42%	0%	85%	26%	1%
Three factor	100%	99%	100%	100%	99%	95%	98%	75%	53%
Four factor	100%	100%	100%	100%	100%	100%	100%	95%	87%
> .95									
One factor	22%	0%	0%	23%	0%	0%	11%	0%	0%
Two factor	87%	0%	0%	78%	0.40%	0%	48%	6%	0%
Three factor	100%	98%	98%	98%	81%	70%	77%	43%	26%
Four factor	100%	100%	100%	100%	99%	96%	94%	76%	63%

Note. High *r*, high factor intercorrelations; Moderate *r*, moderate factor intercorrelations; Low *r*, low factor intercorrelations; > .90, percentage of replications with a TLI value above .90; > .95, percentage of replications with a TLI value above .95.

only moderate, many models were acceptable based on the .90 cutoff.

Correctly specified models almost always appeared acceptable at both the .90 and .95 levels when factor loadings were high. Most overfactored models also predictably fit the data well. However, three-factor models were still more likely to be rejected when factor loadings were smaller, especially when factor intercorrelations were weak. Indeed, when factor loadings were low several three-factor models were rejected even when the more liberal .90 threshold was used. On average, CFI/TLI values were higher when there were fewer items, leading to fewer rejections across models. Although there were fewer rejections of the three-factor model overall when there were only 30 or 15 items, there still tended to be more of these rejections when factor loadings and factor intercorrelations were weaker.

Follow-up analyses

Figures 1 through 8 depict trends in the fit statistics across the models that feature cross loadings or minor factors (complete tables and figures can be found at: osf.io/f2xua). Generally speaking, models were more likely to fit the data adequately or excellently when cross loadings were included in the population models than they were in the initial analyses. However, despite this general tendency to demonstrate better fit, the major trends across fit indices and model conditions described above were still observable. The results based on population models that included minor factors were largely analogous to what was observed in the initial analyses. Thus, the fit

statistics appeared to primarily respond to underfactoring in the major structure while ignoring the omission of minor factors.

Discussion

This Monte Carlo simulation study considered the utility of popular model fit statistics (chi-square, RMSEA, CFI, and TLI) and their commonly applied cutoff values for guiding the assessment of dimensionality in IFA. Generally speaking, results suggest that caution should be employed when relying on conventional fit cutoff rules for evaluating models in the categorical data context of IFA. The primary conceptual utility of fit statistics for IFA should be to alert researchers to potential underfactoring when identifying the optimal factor structure; however, none of the fit statistic cutoffs performed unequivocally well in this regard. Importantly, the trends described below were largely consistent across both simple and more complex population factor structures.

Summary

Fit indices and their standard cutoff scores were most effective at detecting underfactored solutions when analyses were based on continuous data. This is unsurprising given that these thresholds were derived from models based on continuous data, and supports the notion that fit thresholds can be useful when considering the number of factors to retain. The performance of fit statistics and their popular thresholds was less consistent when data were categorical, with thresholds often demonstrating more

sensitivity to variations in study conditions. The CFI/TLI and their traditional cutoff values were the most reliable in detecting underfactoring; the .95 threshold specifically was able to filter out most underfactored models. However, the performance of these indices and their cutoffs was still partly contingent on the strength of the factor loadings, the strength of the factor intercorrelations, and the number of items on the test.

Interestingly, the CFI/TLI cutoffs were more likely to reject models when factor loadings were smaller in magnitude, while models evidenced more statistically significant misfit based on the chi-square test when the factor loadings were larger. This reinforces how certain aspects of the data can exert differential effects on fit statistics and the effectiveness of their cutoffs given the specific attributes of that statistic/index. For example, on one hand it may be more difficult for a model to “perfectly” reproduce the observed matrix when there are many sizeable interrelations between variables (i.e., greater type one error in the chi-square), but on the other hand, when there are many strong interrelations, modeling them even somewhat accurately provides a much bigger improvement over a “baseline” model in which all variables are specified to be orthogonal (whereas small interrelations between variables make the actual model less of an improvement over the baseline, thus the CFI/TLI being more likely to reject models with low factor loadings; Garrido et al., 2016). Practically, this makes it less likely that all fit statistic thresholds will converge on a similar conclusion in analyses with real data, and more difficult to predict why this is the case as the population model parameter values are unknown.

Some trends, however, were consistent across several fit statistics. For instance, models were more likely to demonstrate nonsignificant misfit/appear acceptable according to both the chi-square and CFI/TLI when factor intercorrelations were larger. Indeed, larger correlations between factors suggests less factor distinguishability, and so it is to be expected that underfactored models will fit better when omitted factors overlap strongly with extracted factors (as there is considerable shared information). Also, although the number of indicators per factor did not have any major bearing on the overall conclusions here, in keeping with previous research (West et al., 2012), models in general tended to fit better when there were fewer indicators per factor.

The one exception to this “fewer indicators” pattern was that under certain conditions the RMSEA cutoffs were more likely to reject underfactored models when there were fewer items. This is especially notable because the typical RMSEA cutoffs consistently indicated that underfactored models fit the data excellently. Regardless of the characteristics of the model, all factor solutions

demonstrated adequate, if not excellent, fit to the data by the conventional standards. This suggests that the traditional cutoff values for the RMSEA are likely too liberal for the situations considered here, which is consistent with recent research on model fit in categorical data analyses (e.g., Garrido et al., 2016; Monroe & Cai, 2015; Yang & Xia, 2015). However, these findings do run counter to the recommendation of Barendse and colleagues (2015) to use the RMSEA for determining (the major) factor structure in the context of IFA. Methodological differences may be responsible for this discrepancy. For example, in Barendse and colleague’s study, there were only 12 indicators in the models (and here the RMSEA was less positive when there were fewer indicators), and there was less category asymmetry in certain indicators. Of particular note, however, was that they specifically recommended the RMSEA for selecting the major factor structure in the presence of minor factors. Our follow-up analyses supported this idea in that the RMSEA suggested that the optimal major factor structure of interest fit well when minor factors were omitted; however, it often also suggested that an underfactored major structure fit well. Overall then, the preponderance of the evidence currently seems to suggest that typical RMSEA cutoffs are relatively insensitive to underfactoring, perhaps especially when there are many indicators.

Although fit statistics and their cutoff values were sporadically able to detect underfactoring, they were totally unable to detect overfactoring. This was expected as extracting more factors will more accurately reproduce the observed data matrix. There are, however, practical ramifications of fit statistics’ inherent limitation in signaling overfactoring. Overfitting exploratory factor models, for instance, could result in the dissemination of tests that are characterized by unstable factor structures, and substantively spurious sub-scales (Frazier & Youngstrom, 2007). It is worth noting that overfactored models were more likely to encounter convergence and estimation issues, suggesting that estimation difficulties could help signal an overfactored solution. This is likely to be an unreliable heuristic however.

Overall, given the well-chronicled issues of fit indices and cutoff values with continuous data, it is perhaps not surprising that the fit statistics and cutoffs examined here demonstrated some problems when applied to dichotomous data with varying degrees of asymmetry. It is, however, worth emphasizing that these results do not necessarily imply a problem with the fit statistic *per se*. For example, given the nature of the CFI/TLI (i.e., comparing the current model to a baseline), it is to be expected that they would, and should, look more favorably on underfactored models where there is a higher degree of factor intercorrelation in the population model. Indeed, reinforcing

the sentiments of many ambivalent commentators (e.g., Marsh et al., 2004), the major issue is that the commonly applied cutoff rules, not the statistics themselves, are likely to be misleading under many circumstances.

Implications

The results here demonstrate that fit statistics, or more accurately the common fit cutoffs employed in the literature, are at best imperfect tools for guiding decisions regarding dimensionality when interpreting an IFA. How then should researchers utilize fit statistics in the context of IFA, if at all? The typical CFI/TLI cutoff scores may be useful in helping to protect against underfactoring (especially the more stringent .95 threshold), while the RMSEA should likely be avoided. There are some circumstances in which the RMSEA may prove useful (e.g., Barendse et al., 2015), but the results here and in other recent work (Garrido et al., 2016; Monroe & Cai, 2015; Yang & Xia, 2015) imply that the RMSEA's standard cutoff rules are often far too liberal when working with categorical data. Of course, although the CFI/TLI appeared to be the most practically useful of those indices considered here, the effectiveness of their thresholds was still somewhat contingent on certain data characteristics (e.g., number of indicators, factor intercorrelations, and number of cross loadings).

Caution is therefore warranted when applying fit statistics and their typical thresholds to IFA models, at least when working with dichotomous items. Indeed, decisions regarding dimensionality should likely not be made on the basis of model fit alone. Instead, fit should be used in conjunction with other methods for assessing dimensionality. For example, there is evidence that parallel analysis may be an effective tool in the context of IFA (Garrido et al., 2016). Parallel analysis is not widely available for dichotomous data, and as noted, parallel analysis and other more traditional approaches to assessing dimensionality may not always translate well into the context of IFA (Wirth & Edwards, 2008).

Still, despite the limitations of model fit thresholds in isolation, when paired with other methods (e.g., scree plot analysis), model fit may help to bolster the argument for a given factor solution. That is, to the extent that multiple methods converge on a conceptually coherent solution, it is possible to have greater confidence in that solution, even if there are documented flaws with any one method under certain circumstances. Our findings indicate, however, that if model fit is being used to compliment other methods, prioritizing the CFI/TLI while potentially discounting the RMSEA is warranted (Garrido et al., 2016; Monroe & Cai, 2015; Yang & Xia, 2015).

An alternative approach may be to develop different cutoff values for the RMSEA (and other fit statistics)

specifically for IFA. However, the sensitivity of fit cutoffs to various model and data characteristics demonstrated here and elsewhere makes the development of widely applicable standards infeasible. Garrido and colleagues (2016) recommended that researchers run simulations based on the specific characteristics of their current data (e.g., number of items, number of categories, and sample size) to help identify study-specific fit thresholds. We share this position, and in order to facilitate this approach we have included a detailed program with this report, available on the Open Science Framework (osf.io/f2xua), to aid researchers in conducting their own simulations to examine how fit statistics are like to function in their own IFA analyses. This should help researchers across domains more easily explore the general functioning of fit indices in IFA with scenarios that more closely approximate their actual data.

Limitations and future directions

As is the case in all Monte Carlo studies, there are a number of variables that were not manipulated, and a number of levels of variables that were manipulated that were not considered. For example, the current study relied on only one constant sample size, and only considered dichotomous items. Further, asymmetrical items were mixed together with relatively symmetrical items in a given data set. These data characteristics have “real-world relevance” (e.g., a set of 500 dichotomous responses with some asymmetrical items is a likely scenario when conducting a study on psychopathological symptomatology, or personality, in the general population), but hinder generalizability, and make it more difficult to tease apart certain effects (e.g., would the RMSEA's thresholds have performed so poorly if all items were more symmetrical?). Still, even if the present results are somewhat limited in their scope and applicability, they at least demonstrate that under some conditions standard fit statistic cutoff values will likely suffer from the shortcomings illustrated.

Additionally, it should be noted that many of the population models used here were not especially complex, with three major factors, equal factor loadings, no cross loadings, and repeating sets of thresholds. Given the small body of existing literature on IFA and model fit, and the comparability to the confirmatory factor analytic models commonly used in establishing and evaluating fit cutoffs (e.g., Hu & Bentler, 1999), this represented a reasonable step. Further, additional analyses that considered slightly more complex population models provided results consistent with the initial analyses. Notably, even despite the often clean, favorable setup, the performance of the fit cutoff values was rather mixed. This seems to imply that relying on standard fit cutoff values when examining more

complex data might also be potentially problematic, if not more so. Indeed, fit statistic thresholds were even less likely to detect underfactoring when cross loadings were included.

Finally, the current study did not examine how fit statistics work in tandem with other methods that are typically used to assess dimensionality (e.g., parallel analysis and scree plots). As noted above, it may be the case that fit thresholds are most effective when considered together with these other tools. To be sure, there is evidence that parallel analysis may be effective in IFA (Garrido et al., 2016) which suggests benefits of a more holistic approach. Currently, however, Mplus cannot directly perform parallel analyses with ordinal data and thus parallel analyses were not considered in the present analyses to be consistent with typical practice. This approach may limit the scope of our conclusions, but it maintains the current results' direct applicability to the Mplus modeling framework as it is typically encountered.

Future work should attempt to address these limitations and more thoroughly flesh out the functioning of fit statistics when data are categorical under different conditions. It would also be informative to extend these findings to scenarios in which there are more than two categories. Further, despite the noted difficulties it may be beneficial to attempt to develop adjustments for certain fit statistics (e.g., RMSEA) for when they are applied to categorical data (e.g., Monroe & Cai, 2015). That is, existing fit statistics may need to simply be "fine-tuned" for categorical data. Finally, future work should consider the issue of comparative fit. The present study only examined absolute fit; however, it is common for nested factor structures to be directly tested against one another. Indeed, such tests can offer powerful additional evidence for a certain factor structure. These tests are most commonly performed by comparing the difference in chi-square, but there is also evidence with continuous data that the change in CFI and RMSEA can be used to compare competing measurement models (e.g., Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). The generalizability of these tests to the context of IFA is a fruitful avenue for future investigation

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for

the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grants R305A110293 and R324A150063 from the Institute of Education Sciences, U.S. Department of Education.

Acknowledgments: The authors would like to thank the reviewers for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the Institute of Education Sciences is not intended and should not be inferred.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

- Asparouhov, T., & Muthen, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 397–438.
- Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(2), 211–240. doi:10.1080/105510801922340
- Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), doi:10.1080/10705511.2014.859510
- Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 87–101. doi:10.1080/1075511.2014.934850
- Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, 42, 815–824. doi:10.1016/j.paid.2006.09.018
- Beauducel, A. (2001). On the generalizability of factors: The influence of changing contexts of variables on different methods of factor extraction. *Methods of Psychological Research Online*, 6(1), 69–96.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186–203. doi:10.1207/s15328007sem13022
- Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(1), 41–75. doi:10.1207/s15328007sem12013
- Boivard, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In *Handbook of structural*

- equation modeling (pp. 495–511). New York, NY: The Guilford Press.
- Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, 38(1), 25–56. doi:10.1207/S15327906MBR3801_2
- Brown, T. A. (2006). *Confirmatory factor analysis for applied researchers*. New York, NY: The Guilford Press.
- Brown, T. A. (2013). Latent variable measurement models. In T. A. Little (Ed.), *The Oxford handbook of quantitative methods, volume 2: Statistical analysis* (pp. 257–280). New York, NY: Oxford University Press.
- Browne, M. W., & Cudeck (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: SAGE Publications, Inc.
- Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness of fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, 59, 173–194. doi:10.1348/000711005X66419
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276. doi:10.1207/s15327906mbr0102_10
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. doi:10.1080/10705510701301834
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cut-off points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36(4), 462–494. doi:10.1177/0049124108314720
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255. doi:10.1207/S15328007SEM09025
- Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research*, 32(2), 208–252. doi:10.1177/0049124103256130
- Davey, A. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(4), 578–597. doi:10.1207/s15328007sem1204_4
- DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3), 425–438. doi:10.1080/10705511.2014.915373
- Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, 47, 309–326.
- Edwards, M. C., Wirth, R. J., Houts, C. R., & Xi, N. (2012). Categorical data in the structural equation modeling framework. In *Handbook of structural equation modeling* (pp. 289–311). New York, NY: The Guilford Press.
- Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(3), 343–367. doi:10.1207/s15328007sem12031
- Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, 42(3), 509–529. doi:10.1080/00273170701382864
- Ferrando, P. J., & Loreizno-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicologica*, 21, 301–323.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466–491. doi:10.1037/1082-989X.9.4.466
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(4), 625–641. doi:10.1080/10705510903203573
- Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, 35, 169–182. doi:10.1016/j.intell.2006.07.002
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, 21(1), 93–111. doi:10.1037/met0000064
- Goldberg, L. R., & Velicer, W. F. (2006). Principles of exploratory factor analysis. In S. Strack (Ed.), *Differentiating normal and abnormal personality: Second edition* (pp. 1–31). New York, NY: Springer.
- Hallquist, M., & Wiley, J. (2015). MplusAutomation: Automating Mplus model estimation and interpretation. R package version 0.6-3. Retrieved from <http://CRAN.R-project.org/package=MplusAutomation>
- Hayashi, M. K., Bentler, P. M., & Yuan, K. H. (2007). One the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 505–526. doi:10.1080/10705510701301891
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393–416. doi:10.1177/0013164405282485
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. doi:10.1007/BF02289447
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A multidisciplinary Journal*, 6(1), 1–55. doi:10.1080/10705519909540118
- Kaiser, H. F. (1960). The application of electronic computers for factor analysis. *Educational and Psychological Measurement*, 20, 141–151. doi:10.1177/001316446002000116
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507. doi:10.1177/0049124114543236
- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(3), 333–351. doi:10.1207/S15328007SEM10031

- Kline, P. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.
- Lee, K., & Ashton, M. C. (2007). Factor analysis in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 424–443). New York, NY: The Guilford Press.
- Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, 43, 495–507. doi:10.1007/s11135-007-9133-z
- MacCallum, R. C. (2009). Factor Analysis. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 123–147). London: SAGE Publications Ltd.
- MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51, 201–226. doi:10.1146/annurev.psych.51.1.201
- Markand, D. (2007). The golden rule is that there are no golden rules: A commentary of Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences*, 42, 851–858. doi:10.1016/j.paid.2006.09.023
- Marsh, H. W., Hau, K., Balla, J., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2), 181–220. doi:10.1207/s15327906mbr33021
- Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing.
- McDonald, R. P. (2013). Modern test theory. In T. A. Little (Ed.), *The oxford handbook of quantitative methods, volume 1: Foundations* (pp. 118–143). New York, NY: Oxford University Press.
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3), 568–592. doi:10.101037/0021-9010.93.3.568
- Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, 50(6), 569–583. doi:10.1080/00273171.2015.1032398
- Mulaik, S. (2007). There is a place for approximate fit in structural equation modeling. *Personality and Individual Differences*, 42, 883–891. doi:10.1016/j.paid.2006.10.024
- Muthen, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: SAGE Publications, Inc.
- Muthen, B. O., du Toit, S. H. C., & Spisic (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes (Report No. 75). Location: www.statmodel.com
- Muthen, L. K., & Muthen, B. O. (1998–2015). *Mplus user's guide*. Seventh Edition. Los Angeles, CA: Muthen & Muthen.
- Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *The Journal of Experimental Education*, 68(3), 251–268. doi:10.1080/00220970009600095
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2(1), 13–43. doi:10.1207/S15328031US0201_02
- Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48(1), 28–56. doi:10.1080/00273171.2012.710386
- R Core Team. (2016). *R: A language and environment for statistical computing*. (3.3.1) [computer program]. Vienna, Austria: R Foundation for Statistical Computing.
- Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.) *Handbook of item response theory modeling* (pp. 13–40). New York, NY: Taylor & Francis.
- Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. doi:10.1037/a0029315
- Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, 28(12), 1629–1646. doi:10.1177/014616702237645
- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, 16(4), 561–582. doi:10.1080/10705510903203433
- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 167–180. doi:10.1080/10705511.2014.882658
- Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Education and Psychological Measurement*, 72(6), 910–932. doi:10.1177/0013164412452564
- Shin, H. C. (2013). Weighted least squares estimation with sampling weights. *Journal of Mathematics and Statistics*.
- Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42, 893–898. doi:10.1016/j.paid.2006.09.017
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In *Handbook of structural equation modeling* (pp. 209–231). New York, NY: The Guilford Press.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58–70. doi:10.37/1082-989X.12.1.58
- Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: SAGE Publications.
- Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavioral Research*, 47, 756–772. doi:10.3758/s13428-014-0499-2
- Yuan, K. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, 40(1), 115–148. doi:10.1207/s15327906mbr4001